EDGEWOOD CHEMICAL BIOLOGICAL CENTER
U.S. ARMY RESEARCH, DEVELOPMENT AND ENGINEERING COMMAND
Aberdeen Proving Ground, MD 21010-5424

ECBC-TR-1125

SCIENCE OF DECISION MAKING: A DATA-MODELING APPROACH

Samir V. Deshpande

SCIENCE AND TECHNOLOGY CORPORATION
Edgewood, MD 21040-2734

Rabih E. Jabbour

RESEARCH AND TECHNOLOGY DIRECTORATE

October 2013

Disclaimer

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorizing documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
XX-10-2013	Final	Sep 2011 - Oct 2012

4. TITLE AND SUBTITLE
Science of Decision Making: A Data-Modeling Approach

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Deshpande, Samir V. (STC); and Jabbour, Rabih E. (ECBC)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Science and Technology Corporation, 500 Edgewood Road, Suite 205, Edgewood, MD 21040-2734
Director, ECBC, ATTN: RDCB-DRD-D, APG, MD 21010-5424

8. PERFORMING ORGANIZATION REPORT NUMBER
ECBC-TR-1125

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Edgewood Chemical Biological Center, Seedling Program, APG, MD 21010-5424

10. SPONSOR/MONITOR'S ACRONYM(S)
ECBC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
We have developed a parallel data analysis algorithm for peptide classification, which is used for microbial identification. This algorithm was based on data generated from the commercially available algorithms SEQUEST and OMSSA. The outputs from those algorithms were analyzed to determine a probability score for the identified peptides and their associated proteins. The statistical analyses and data interpretation using our proposed approach showed that we can lower the false-discovery rate by using common proteins from both algorithms. This approach showed that the identification accuracy and reliable classification of microbes were improved without increasing the data analysis time. In summary, we have a higher confidence in the identification process and a reduced bottleneck in data analysis through the use of the new algorithm.

15. SUBJECT TERMS
Microbial identification; Proteomics; Bioinformatics; Algorithm
Protein identification; Database; Data modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Renu B. Rastogi
U	U	U	UU	20	19b. TELEPHONE NUMBER (include area code) (410) 436-7545

Blank

PREFACE

The work described in this report was authorized under the U.S. Army Edgewood Chemical Biological Center (Aberdeen Proving Ground, MD) In-House Laboratory Independent Research Program. This work was started in September 2011 and completed in October 2012.

The use of either trade or manufacturers' names in this report does not constitute an official endorsement of any commercial products. This report may not be cited for purposes of advertisement.

This report has been approved for public release.

Acknowledgments

The authors would like to thank the Seedling program committee for their financial support of this work, and to extend special thanks to Patrick McCubbin (OptiMetrics) for his help in the editing of this report.

Blank

CONTENTS

FIGURES

TABLES

SCIENCE OF DECISION MAKING: A DATA-MODELING APPROACH

1. INTRODUCTION

Peptide mass fingerprinting (PMF)-based identification algorithms using mass spectrometry (MS) data were developed in the early 1990s. One of the early PMF algorithms, SEQUEST (Yates Lab, The Scripps Research Institute; La Jolla, CA), was widely used by the scientific community to decipher the sequence information of peptides generated from tandem MS analysis. This software is commercially available and solely distributed by Thermo Fisher Scientific (Tewksbury, MA) (*1*,*2*). Later, other PMF algorithms were developed and reported in literature. These included Mascot from Matrix Science, Inc. (Boston, MA) and open-source software such as the open mass spectrometry search algorithm (OMSSA) from the National Center for Biotechnology Information (Bethesda, MD) and X!Tandem from The Global Proteome Machine Organization, which is an online database (*3*–*5*). These algorithms assign a peptide sequence, along with a matching score of the experimental ion product mass spectrum, to a theoretical ion product mass that is derived from the protein sequences in a given proteome database. The resulting peptide-spectrum match (PSM) score is computed by either *descriptive*, *interpretative*, *stochastic*, or *probabilistic* modeling methods and are used to provide discrimination between true-positive (TP) and false-positive (FP) peptide identification (*6*). The aforementioned PMF algorithms have an inherent extensive computational time requirement that becomes cumbersome for high-throughput proteomic analysis. Therefore, there is a need to overcome such obstacles in data analysis procedure throughout the development of relatively rapid tools that are capable of identification and classification of microbes in near realtime settings.

The PMF algorithms create overlap between TP and FP peptide identifications (*7*), whereby the identified FP peptides lower the overall confidence for the identified TP peptides. Keller, et al. developed the PSM-scoring algorithm on the basis of machine learning methods such as linear discriminant analysis (*8*). This algorithm provides a higher confidence level in peptide identification because each spectrum is discriminated by weighing each vector feature and by providing a relative weight to that peptide MS spectrum. Many researchers use a decoy database that contains reversed protein sequences to score the FP peptide identification and, thereby, compute a false-discovery rate (FDR) score (*9*). This decoy database has two limitations: (a) the database search time is doubled, and (b) a suitable decoy database cannot be generated for all applications, especially when the researcher is not doing a targeted database search (*10*).

Gupta, et al. (2011) stated “that target-decoy approach (TDA) is not needed when accurate p-values of individual peptide-spectrum matches are available” (*11*). Moreover, when using a decoy database it is difficult to maintain the mass and amino acid composition of the target and decoy peptides.

To overcome the issues of over-fitting a vector feature of the spectrum and the use of a decoy database, and to lower dynamically the FDR score, we have developed a parallel data analysis algorithm called "Merlin". Merlin can be used to analyze PSM results from the SEQUEST and OMSSA algorithms, and it has the potential to analyze results from other PMF algorithms as well.

The Merlin algorithm employs multiple scores such as cross-correlation (X_{Corr}), preliminary score (S_p), and mass differences coefficient (ΔCn) from SEQUEST. Merlin also incorporates the probability value (*p*-value) from PeptideProphet and the expected value (*E*-value) from OMSSA to compute the most probable PSM for the identification and classification of an organism in the analyzed experimental sample. (The *E*-value is a parameter that describes the number of successes expected when searching a database of a particular size.)

In the future, we plan to incorporate other open- and closed-source PMF algorithms to provide a robust and automated PMF algorithm such as Merlin, which would be capable of improving the confidence score of identified peptides during the proteomics data processing. This algorithm could be integrated within the U.S. Army Edgewood Chemical Biological Center in-house-developed microbial identification tool, ABOid (*12*).

2. METHODS

2.1 *Escherichia coli* Strain O157:H7 Sample Preparation

The *E. coli* strain O157:H7 was grown in trypticase soy broth, contained in an orbital shaker (125 rpm) at 37 °C, until the bacteria reached the late exponential phase (~10^8 cfu/mL). The cell culture was stored at 4 °C until it reached fractionation. To isolate the secreted protein fractions, 30 mL of culture was centrifuged at 11,300×*g* for 1 h using a Beckman J2-MC centrifuge (Indianapolis, IN). The supernatant was decanted to separate it from the pellet. This supernatant, which contains the secreted proteins, is referred to as the *secreted fraction*. The pellet was resuspended in ~3.5 mL of 100 mM ammonium bicarbonate (ABC). This extracellular suspension was divided into three aliquots of approximately equal volume. The cell pellet extracellular samples were thawed and lysed by ultrasonication (25 s on and 5 s off for a total of 4 min) using a Branson Digital Sonifier (Danbury, CT). The lysate was centrifuged at 14,000 rpm for 20 min at 10 °C using a Beckman GS-15R centrifuge. Samples were frozen at –25 °C for up to 4 days.

2.2 Bacterial Sample Processing

Samples were prepared for liquid chromatography (LC) tandem MS (LC–MS/MS) in a similar manner to that previously reported by Jabbour et al. (*13*). Proteins were extracted from the secreted fractions by transferring each sample to a separate Microcon YM-3 filter unit (Millipore, Billerica, MA) and centrifuging the samples at 14,100×*g* for 20–30 min. The filters were each centrifuged three times at 14,000×*g* for 25 min, with a 200 µL ABC wash between centrifugations.

The proteins in the retentate were denatured at 40 °C for 1 h with 270 µL of 7.2 M urea and 30 µg/mL dithiothreitol in ABC. The urea was removed by centrifugation at 14,100×*g* for 30–40 min. The retentate was washed three times using 150 µL ABC, followed by centrifugation at 14,100×*g* for 30–40 min using an Eppendorf centrifuge 5415C or 5415D (Eppendorf North America; Westbury, NY). The filter unit was then transferred to a new receptor tube and the proteins in the retentate were digested overnight at 37 °C with 5 µL sequencing-grade trypsin (Product No. 511A; Promega; Madison, WI) in 10 µL acetonitrile and 240 µL ABC. The tryptic peptides were isolated by centrifuging at 14,100×*g* for 20–30 min.

2.3 LC–MS/MS Experiments

In a manner similar to that previously described by Jabbour et al. (*13*), the tryptic peptides were separated on a capillary column using the Dionex UltiMate 3000 (Sunnyvale, CA). The resolved peptides were then sprayed into a linear ion trap MS (LTQ XL; Thermo Scientific; San Jose, CA). The product ion mass spectra were obtained using the data-dependent acquisition mode with a survey scan, followed by performing an MS/MS evaluation on the top five most-intense precursor ions.

3. RESULTS AND DISCUSSION

3.1 Database Search and Data Analysis

A proteome database was constructed in a FASTA format derived from the *E. coli* O157:H7 strain Sakai genome obtained from the National Center for Biotechnology Information (NCBI) genomic database repository (http://www.ncbi.nlm.nih.gov, accessed August 14, 2012). The constructed proteome database included 115 protein sequences from all potential laboratory contaminants. The constructed proteome database also consisted of 5433 proteins that were used in this study. The targeted proteins in the proteome database were in silico digested using trypsin to perform enzymatic cleavage and to obtain the theoretical product ion spectra of all potential peptides. Then, the proteome database was indexed in FASTA format for compatibility with the examined algorithms (SEQUEST and OMSSA) listed in Table 1.

The experimental product ion spectra in *.RAW file format were obtained using the LTQ XL MS and converted into the mass-to-charge extensible markup language (mzXML) format using a file-conversion tool developed by Seattle Proteome Center at the Institute of System Biology (Seattle, WA) (*14*). Three replicate suspension samples analyzed on tandem MS/MS that resulted in 43501 MS/MS spectra were searched against the constructed proteome database according to the parameters listed in Table 1 and including two additional parameters: (a) mass tolerance of 2.50000 amu and (b) fragment ion tolerance of 1.00000 amu.

Table 1. Protein Sequence In Silico Digestion Parameters

Parameter	Value
FASTA database	EC_Sakai.fasta
FASTA index	EC_Sakai.fasta.idx
FASTA digest	EC_Sakai.fasta.dgt
Enzyme name	Trypsin (KR)
Mass range	600–3500 m/z
Sequence length	5–35
Mass type	Monoisotopic
Missed cleavage sites	2

The MS files (.RAW files) were submitted to SEQUEST. The same files that were converted into mzXML files were submitted to OMSSA for database searching. The output from the SEQUEST database, without any threshold cutoff, were submitted to the ABOid software to derive the probability score of peptides production ion spectra and then converted into a comma-separated file format (.csv) to concatenate the spectral files into one file. The PSM results from the OMSSA database were also exported to a .csv file format. The .csv files generated from SEQUEST contain information such as the scan number, peptide, X_{Corr}, S_p, ΔC_n, RSp (rank score), M+H (molecular ion), protein name, and accession number. The .csv files generated by the OMSSA database contains information like scan number, peptide sequence, peptide mass, protein name and mass, accession number, *E*-value, and *p*-value.

For each analyzed sample, the .csv files resulting from the OMSSA and SEQUEST algorithms were also submitted through the Merlin algorithm to extract the common proteins identified previously. The common proteins, their weighing factors, and database parameters were submitted again to the ABOid software for identification and computation of the probability score. Peptide sequences with probability scores of 95% and higher were retained and used to generate a binary matrix of sequence-to-bacterium assignments. The binary matrix was populated by matching the peptides with corresponding proteins in the constructed proteome database and assigning a score of one for a match and zero for a mismatch. The columns in the binary matrix represent the proteome of bacteria and contaminants in the database, the rows represent identified tryptic peptide sequences that were obtained from tandem MS spectral processing. A sample microorganism is matched with a database bacterium by the number of unique peptides that remained after filtering of the degenerate peptides from the binary matrix. Verification of the classification and identification of candidate microorganisms is performed through hierarchical clustering analysis and taxonomic classification as shown by the ABOid software. The flowchart for the Merlin process is shown in Figure 1.

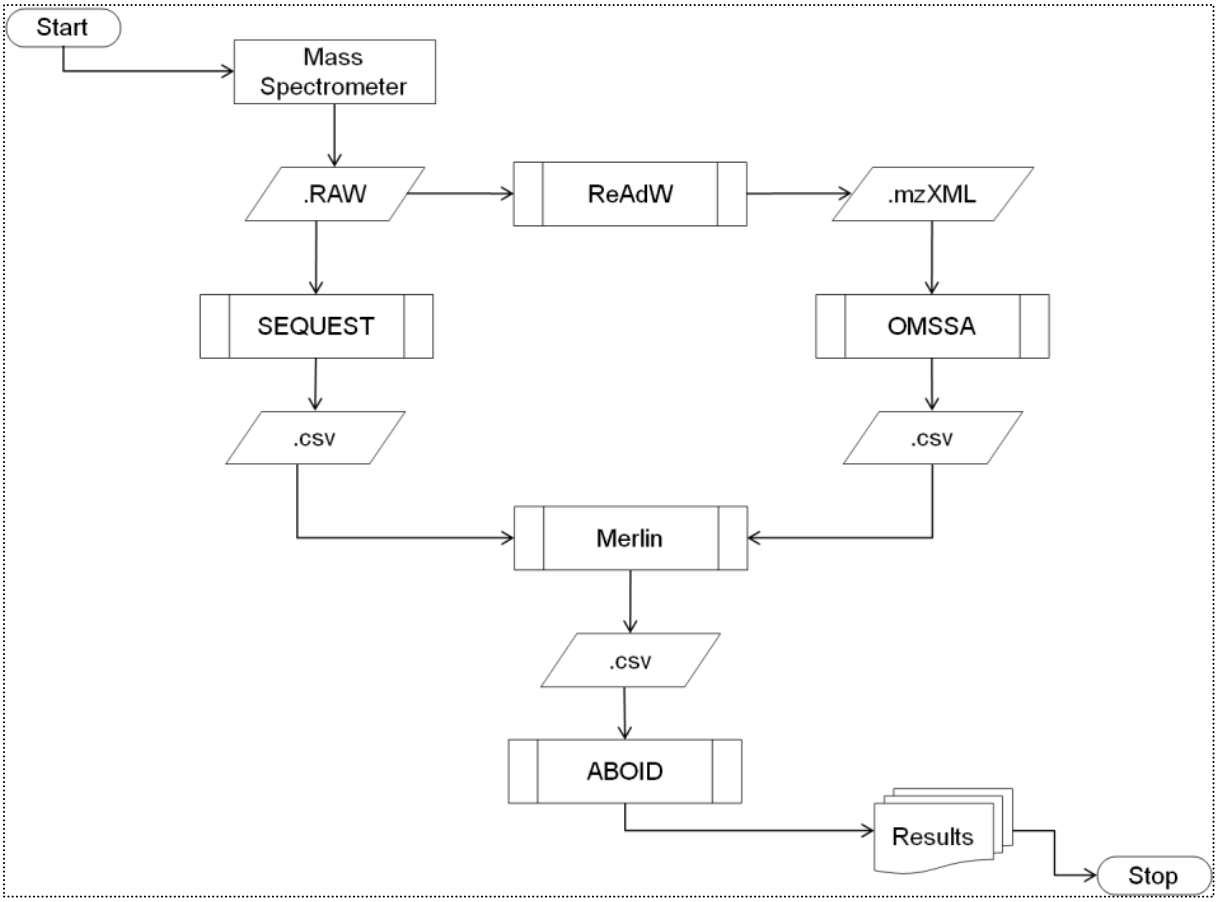

Figure 1. Flow chart for Merlin algorithm.

Table 2 shows the total number of unique proteins observed from the two database-searching algorithms and the common proteins identified by the Merlin algorithm. Although the number of common proteins identified using Merlin was relatively lower than that of the other algorithms, the identification score for the bacteria was higher using the protein list from Merlin than that of the other algorithms.

Table 2. Database-Searching Comparison and Number of Unique Proteins Observed

Sample ID	Spectra	SEQUEST	OMSSA	Merlin
2011-02-10-02	13265	103	107	80
2011-02-10-03	14728	82	92	68
2011-02-10-05	15508	119	140	101

Figure 2 shows a Venn diagram of the analyzed bacterial samples and the number of candidate proteins identified using the Merlin algorithm. The results showed an increase of common proteins in replicate analyses with Merlin, which was lower using the other algorithms individually.

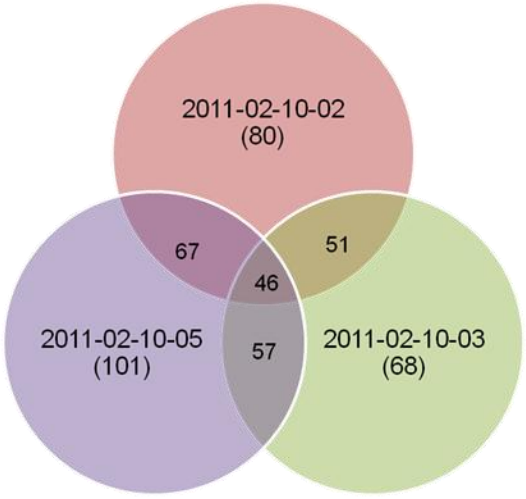

Figure 2. Venn diagram for Merlin results obtained from the replicate bacterial samples.

4. CONCLUSIONS

This study showed that the use of a single PMF algorithm could result in a higher FDR value when compared with a combinatorial approach that concurrently retains spectral information from diverse individual algorithms. This conclusion was based on statistical confidence using Bayesian and Gaussian PMF algorithms to lower the FP rate and eliminate the data analysis bottleneck.

Additional studies are needed to incorporate the de novo analysis and identify the best-fitting peptides without using database-searching tools. The Poisson distribution should be used to match de novo output with peptides identified by database-searching tools. In addition, the incorporation of a receiver-operating characteristics curve will enable computation of the probability cutoff value for analysis. Such studies will expand the algorithms to provide enhanced selectivity.

LITERATURE CITED

1. Eng, J.; McCormack, A.; Yates, J. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1999**, *5*(11), 976–989.

2. Thermo Scientific Home Page. http://www.thermoscientific.com (accessed on 10/09/12).

3. Perkins, D.N.; Pappin, D.J.; Creasy, D.M.; Cottrell, J.S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999**, *20*, 3551–3567.

4. Geer, L.Y.; Markey, S.P.; Kowalak, J.A.; Wagner, J.; Xu, M.; Maynard, D.M.; Yang, X.; Shi, W.; Bryant, S.H. Open Mass Spectrometry Search Algorithm. *J. Proteome Res.* **2004**, *3*(5), 958–964.

5. Craig, R.; Beavis R.C. TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatic* **2004**, *20*(9), 1466–1467.

6. Sadygov, R.G., Cociorva, D.; Yates, J.R. Large-Scale Database Searching Using Tandem Mass Spectra: Looking Up the Answer in the Back of the Book. *Nature Methods* **2004**, *1*(3), 195–202.

7. MacCoss, M.J.; Wu, C.C.; Yates, J.R. Probability-Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm. *Anal. Chem.* **2002**, *74*, 5593–5599.

8. Keller, A.; Nesvizhskii, A.I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem*. **2002**, *74*(20), 5383–5392.

9. Elias, J.E.; Gibbons, F.D.; King, O.D.; Roth, F.P.; Gygi, S.P. Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nature Biotechnology* **2004**, *22*, 214–219.

10. Kim S.; Gupta N.; Pevzner P.A. Spectral Probabilities and Generating Functions of Tandem Mass Spectra: A Strike against Decoy Databases. *J. Proteome Res*. **2008**, *7*(8), 3354–3363.

11. Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P.A. Target-Decoy Approach and False Discovery Rate: When Things May Go Wrong. *J. Am. Soc. Mass Spectrom*. **2011**, *22*(7), 1111–1120.

12. Deshpande, S.V.; Jabbour, R.E.; Snyder P.A.; Stanford, M.; Wick, C.H.; Zulich, A.W. ABOid: A Software for Automated Identification and Phyloproteomics Classification of Tandem Mass Spectrometric Data. *J. Chromatogr. Sep. Tech.* **2011**, S5:001.

13. Jabbour, R.E.; Deshpande, S.V.; Wade, M.M.; Stanford, M.F.; Wick, C.H.; Zulich, A.W.; Skowronski, E.W.; Snyder, A.P. Double-Blind Characterization of Non-Genome-Sequenced Bacteria by Mass Spectrometry-Based Proteomics. *Appl. Environ. Microbiol*. **2010**, *76*(11), 3637–3644.

14. Pedrioli, P.G.A.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E.W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R.H.; Apweiler, R.; Cheung, K.; Costello, C.E.; Hermjakob, H.; Julian, R.K.; Kapp, E.; McComb, M.E.; Oliver, S.G.; Omenn, G.; Paton, N.W.; Simpson, R.; Smith, R.; Taylor, C.F.; Zhu, W.; Aebersold, R. A Common Open Representation of Mass Spectrometry Data and its Application to Proteomics Research. *Nat. Biotechnol*. **2004**, *22*(11), 1459–1466.

ACRONYMS AND ABBREVIATIONS

ΔCn	mass differences coefficient
ABC	ammonium bicarbonate
E-value	expected value
FDR	false-discovery fate
FP	false-positive (peptide identification)
LC	liquid chromatography
LC–MS/MS	liquid chromatography–tandem mass spectrometry
MS	mass spectrometry
mzXML	mass-to-charge extensible markup language
OMSSA	open mass spectrometry search algorithm
PMF	peptide mass fingerprinting
PSM	peptide-spectrum match
p-value	probability value
RSp	rank score
S_p	preliminary score
TP	true-positive (peptide identification)
X_{Corr}	cross-correlation


EDGEWOOD CHEMICAL BIOLOGICAL CENTER
A U.S. Army RDECOM Laboratory